ALY6000

Probability Theory and Introductory Statistics

Module 5 – R Practice 5

Name: Duc Viet Mai

Student ID: 002661113

Date: December 9, 2022

I. INTRODUCTION

In this paper, I will establish a correlation and regression table for the data set about testing the effect of Gabapentin drugs based on specific traits of patients. The first part includes the correlation table of five variables (smoking behaviors, hormone replacement therapy, white blood cells, age, and outcome). And the second part contains a regression table to test the relationship between the dependent and independent variables.

II. DATA ANALYSIS

1. Part 1

```
##                 smoker   hrt_months         wbc         age     outcome
## smoker      1.000000000 -0.006440119  0.17798414 -0.12818466  0.01510919
## hrt_months -0.006440119  1.000000000  0.14778555  0.16682702  0.02223168
## wbc         0.177984137  0.147785546  1.00000000  0.05343805 -0.03683491
## age        -0.128184664  0.166827025  0.05343805  1.00000000  0.09041397
## outcome     0.015109187  0.022231678 -0.03683491  0.09041397  1.00000000
```

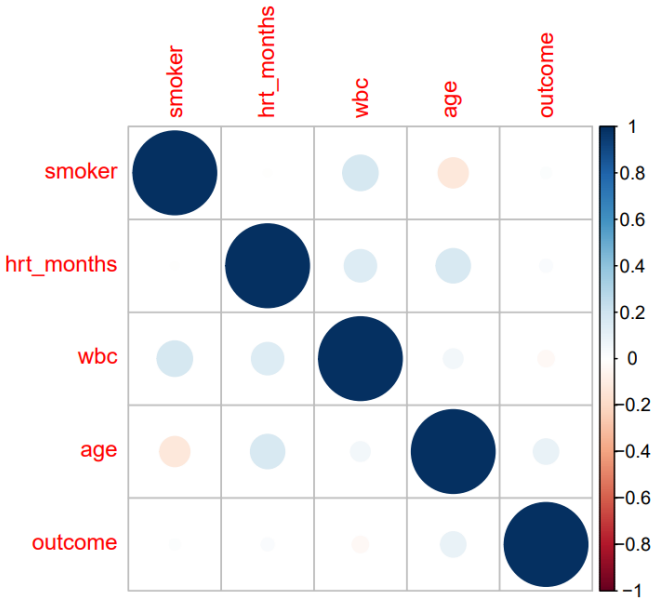
Starting with smoking behavior, the correlation between it and hormone replacement therapy is nearly 0 (-0.006), or no correlation. Next, the correlation between smoking behavior and white blood cells is 0.17, which is a negligible relationship, but we know it is positive. The relationship might be understandable since smoking tends to increase the white blood cells in the human body, resulting in stroke diseases. With age, the correlation value is -0.13, which is also not taken into account; However, I aware it is a negative relationship. As common sense, smoking will indeed decrease our life span. Finally, with the outcome, the correlation value is 0.015, which is also nearly no correlation between it and smoking.

Then, we look at hormone replacement therapy, as the relationship between it and white blood cells (0.15), age (0.17), and outcome (0.02) are all negligible relationships between 0 and 0.3. Yet, because they are positive relationships, these results mean that these variables will increase and decrease side by side.

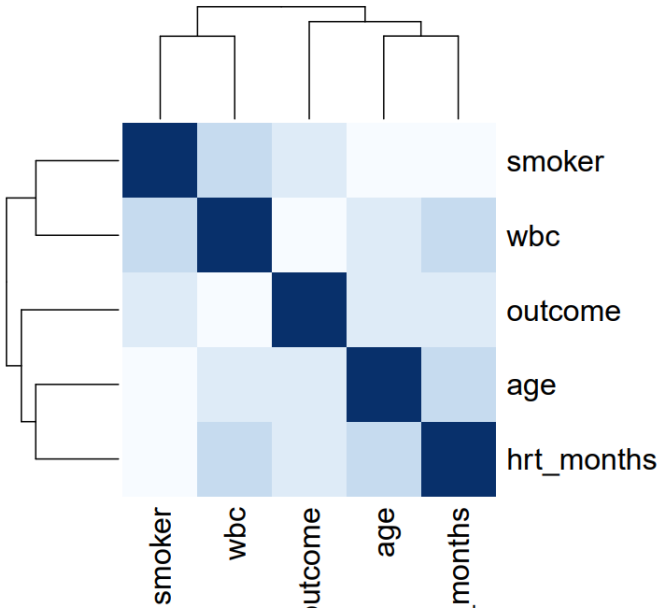
Next, the relationship between white blood cells and age is insignificant (positive) since the correlation value is 0.05. In contrast, the relation is also trivial (negative), with a correlation value of -0.03. And lastly, the relationship between age and outcome is no correlation as the correlation value is 0.09.

The reason behind zero correlation between most variables is that we use Pearson correlation, which focuses mainly on linear relationships or one-to-one change. Variables in our data set might have a different rate of change, meaning that one unit change in this

variable relates to less/more than one unit change in others. Consequently, if we plot them out, we might notice a curve-linear relationship. In this case, it is better to use Spearman/Kendall correlation, which measures the monotonic relationship. Monotonic means that variables decrease and increase in the same direction but not necessarily in the same proportion (one-to-one).



Because the correlation chart is diagnostic, you can receive the same correlation value if you flip the chart around the diagonal axis. I also visualize the chart with a heatmap so you can quickly imagine the correlation value and the strength of the relationship between variables. The lighter the color, the weaker the relationship, and vice versa.



2. Part 2

```
##
## Call:
## lm(formula = outcome ~ smoker + hrt_months + wbc + age, data = newdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.441  -8.175  -3.109   4.036 134.518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.854293   4.901009   0.378  0.70526
## smoker       1.442848   1.311110   1.100  0.27140
## hrt_months   0.004107   0.009686   0.424  0.67169
## wbc         -0.369134   0.242714  -1.521  0.12863
## age          0.258112   0.089309   2.890  0.00394 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.14 on 951 degrees of freedom
## Multiple R-squared:  0.01135,    Adjusted R-squared:  0.007192
## F-statistic:  2.73 on 4 and 951 DF,  p-value: 0.02809
```

We use the correlation table above to check for the relationship between variables and their movement. On the other hand, we exploit the regression table to check for the cause and effect between variables, or in other words, this is for prediction purposes. In the correlation table, we can interchange the data point, unlike in a regression table, where we can not because we need to define which are independent and dependent variables.

In this regression table, we have an intercept value of 1.85. And as noticeable, my dependent variable depends positively on smoking behaviors. Specifically, for one unit change in smoking behaviors, the outcome will increase or decrease by 1.44. In contrast, one unit change in hormone replacement therapy does not seem to change much of the dependent variable since we have a coefficient of nearly 0 (0.004). Next, white blood cells have a negative coefficient of -0.40, which means the higher the white cells in the body, the lower the outcome, and vice versa. Finally, one unit change in age will cause the dependent variable to increase or decrease by 0.26 accordingly.

In addition, the high residual value (12.14 - observations deviate too much from the regression fitted line) and low R-squared value (0.01) mean this regression model does not predict well. Or, we might say that it can only forecast 1% of the observations.

REFERENCES

[1] Bluman, A. G. (2018). *Elementary statistics: A step by step approach.* McGraw-Hill Education. 10th ed, 1- 368. ISBN: 978-1-259-75533-0.

[2] Facer, C. (n.d.). How to create a correlation matrix in R. Display R Blog. Retrieve from: https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/

[3] STHDA (Statistical tools for high-throughput data analysis). (n.d.). *Correlation matrix: A quick start guide to analyze, format and visualize a correlation matrix using R software*. R & Data Mining. Retrieve from: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software